Spatial dialectics

Spatial Dialectics: Pursuing Geospatial Imaginaries with Word Embedding Models and Mapping

By Anouk Lang

The relationship between human beings and their environment is one of the key problematics explored in twentieth-century literature. As modernist studies has turned its attention to contexts beyond Britain, Europe and the United States, so questions around space, place and geography have been necessarily reconfigured to take account of the effects of imperialism and globalization, and to destabilize the Anglo- and Eurocentrism of prevailing critical perspectives on space within modernist writing. Roughly concomitant with the development of these geomodernist approaches, significant advances have been made within the field of spatial humanities by scholars who have sought ways to use powerful GIS software in pursuit of research questions specific to the humanities.[1] Some of the most interesting research in this area has sought to directly confront the difficulties of using software that requires quantitative input to account for the complexities of spatial imaginaries, understood here as an imbricating set of discursive constructs concerned with the elaboration of spatial meanings. While such discursive constructs can sometimes be anchored to locations in the material world with specific latitude and longitude coordinates, they are more likely to occupy an ambiguous position in relation to the exigencies of georeferencing, or even to float entirely free from such constraints. Unravelling the workings of spatial imaginaries within a corpus that combines both georeferenceable and non-georeferenceable entities thus engages one of the core debates animating work in the digital humanities: using technologies that often mandate binary distinctions and discrete categories to represent and interrogate a world of non-binary human experiences.

In this article, I seek to understand how such spatial imaginaries are operating within *The Western Home Monthly* (1901–32), a Canadian household magazine whose publication out of Winnipeg rather than the more populous eastern cities of Canada (where publishers were larger and more established) makes it a particularly interesting periodical to use to investigate place (fig. 1).[2] I suggest ways in which different modes of computational analysis might be used in combination to understand how *The Western Home Monthly* was representing place to its readers, not only in relation to north American, European, and international space, but also in terms of less easily specified geographies around the relationship between humans and the natural world, the topographical variations across the Canadian landmass, and the interplay between historical and contemporary occupation of the land. The aim here is twofold: to understand *The Western Home Monthly* better through this methodological approach, and to evaluate how tractable its digitized version is when methods from the digital humanities are used to investigate the spatial imaginaries that it was engaged in constructing, mediating, and reflecting back to its readership.

[INSERT FIG. 1 ABOUT HERE]

Hannah McGregor and Nicholas van Orden give a full description of this periodical and the digital artefact that represents almost the entirety of its run (24,170 pages and 33,099,536 words), along with details of the material history of its digitization, information about the metadata and encoding schemas used to create it, and a persuasive rationale for situating it as a middlebrow periodical.[3] Their study demonstrates the value of combining close and distant reading modes on a digitized periodical, by showing how topic modeling is able to identify the ways that lexical

patterning around new media such as radios appeared in conjunction both with advertisements and with serialized narratives in which radios featured as indexes to modernity. Here, I apply a different method from machine learning to the *WHM*: word embedding models.[4]

Word Embedding Models and Vector Space

Although word embedding models have been in use within computer science and linguistics for some decades, they are one of the newer methods of text analysis within the digital humanities. Word embedding models are mathematical constructs in which the relationship of every word in the corpus to every other word in the corpus can be mathematically specified in high-dimensional space as a vector, producing what is known as a vector space model. Such models represent each word as "a point in high-dimensional space, where each dimension stands for a context item, and a word's coordinates represent its context counts."[5] This high-dimensional space is commonly reduced down from some tens of thousands of dimensions to a few hundreds—in this study to two hundred dimensions—so that calculations about the proximity of particular words to other words can be more efficiently calculated. In order to represent these relationships on the page or the screen, the model's dimensionality is reduced still further to two or three dimensions using techniques such as t-SNE (t-distributed stochastic neighbor embedding), (fig. 2).[6]

[INSERT FIG. 2 ABOUT HERE]

How do we read an image such as this, whose grand claim of representing the relationality of many thousands of words with a high degree of quantitative precision is at odds with its visual

incoherence? If words appear in spatial proximity to other words in these reduced-dimensionality representations, this means that they occur in the corpus in similar contexts, and are therefore more likely to be similar in their meaning than words that are further apart.[7] The question of what exactly is meant by "similar" here is a key question, and is very much a live one within the fields of computational linguistics and natural language processing. Similarity can for instance denote synonymy, or indicate words which can fill the same syntactic slot (e.g., the cluster of personal pronouns—*i me my you your*—visible in figure 2). As we will go on to see, there are many other ways to interpret proximity between groups of words in a word embedding model, and it is one of my aims here to interrogate this idea of similarity, with reference to the ways that the conceptual convergences and divergences suggested by the model throw light on the ways *The Western Home Monthly* imagines the spaces of Canada in the early decades of the twentieth century. It is my contention that literary critical understandings of *discursive* similarity, articulated in this periodical in relation to discourses of place, may be useful in demonstrating the utility of vector space models when they are used to investigate literary and historical research questions. Another aim is to demonstrate how word embedding models may be useful at different points along the continuum from close reading to so-called distant reading. At the macroanalytic scale, they allow researchers to see the broad brushstrokes of a corpus, but also to move from these broad patterns to investigate mid-level queries about, say, the operation of discourses, and from there to more specific queries about analogies, words which appear only a handful of times in the corpus and so on. In what follows, I seek to give examples of these different levels, and to show what can be learnt from moving back and forth between them. Finally, I suggest some ways in which the results derived from word embedding models might be

combined with GIS mapping in ways that account for both semantic relationships and spatial/geographical relationships.

Word Embedding Models and Gendered Discourse

The power of word embedding models to discern discursive patterning in language is most readily illustrated with gendered discourse, so it is worth beginning with an example of how this is done. Figure 2 gives a two-dimensional rendering of some of the terms in the model, providing a visual analogue for the way they are clustering, but it is possible to obtain much more precise information which quantifies the extent to which terms are similar. Figure 3 shows the words in the *WHM* which are most heavily skewed along the gender binary, i.e., the vectors whose cosine similarities are closest to the cosine similarities of the vector *she* and furthest from the vector *he* (orange lines), and vice versa (green lines).

[INSERT FIG. 3 ABOUT HERE]

The gendered nature of this list is clear. Words designated by orange bars in the upper part of the plot provide stereotypical descriptions of women (*motherly*, *dimpled*, *matronly*, *prettier_than*, *weeping*, *cuddly*), alongside several terms which mark transgressions from approved gender norms (*untidy*, *dowdy*). The roles described center around familial relations and the domestic sphere (*maid*, *aunt*, *aunty*, *elder_sister*, *grandmother*, *mother*, *girls*, *maids*, *mistress*, *seamstress*, *grandma*), while the presence of several variants on the word *husband* in the list rather than *man*, or indeed *father* or *brother*, signal the type of man that appears most frequently near words whose vectors are similar to *she*. Other nouns reinforce these associations (*household tasks*,

trousseau, *rag_doll*, *flushed_cheeks*, *doll*, *apron_strings*, *wedding gown*, *hairpins*) and give a picture of a very clear semantic field of family, home and marriage.

The words in the corpus designated by green bars as being the mostly strongly gendered as male (whose similarity to *he* and distance from *she* are the most pronounced) also belong to a field of similarly stereotypically masculine pursuits: hunting, pioneering, and gambling (*faro* being a gambling game). The words gendered male are also far fewer in number than those gendered female, which is an indication of the fact that more of the words in the corpus cluster closer to *she* than to *he*). This is perhaps not surprising when considering what the advertisements in the *WHM* suggest about its readership: many of them are oriented towards activities that were seen as the province of women during the period of the magazine's run, such as keeping the house clean, and purchasing products such as clothing and toiletries. The short fiction pieces in the magazine are also important in constructing this gendered edifice, as can be seen by the presence of many character names in the list in figure 3.

Exploring the operation of gendered discourse in the *WHM* would be a study of considerable interest in its own right, but while it is easy to pick two vectors to denote the two ends of the gender binary (*he* and *she*), it is not so obvious which terms to choose to facilitate an exploration of a spatial imaginary. I next turn to the question of how a word embedding model might be used to investigate "ways of talking about place": the discursive construction of a spatial imaginary within this periodical.

Discourses of Place and Vector Space Geosemantics

Benedict Anderson's notion of imagined communities has been influential in demonstrating the importance of newspapers in disseminating ideologies such as nationalism across a geographically dispersed population.[8] Although Anderson's argument holds for print culture artefacts across a range of media—magazines as well as newspapers, and middlebrow as well as popular and news publications—the ready availability of digitized newspaper archives have led to the relative neglect of other serialized genres in studies of nationalist discourse. The opportunity to carry out distant reading on an archive close to the entire run of a middlebrow periodical such as the *WHM* is thus a compelling one, as its *difference* from newspapers—the relative proportion of short fiction compared to non-fiction prose, for instance—presents an alternative set of opportunities to observe discursive formations in process. Using word embedding models allows us to investigate the words and concepts which are *analogically* or *discursively* proximate to a particular place name, but not necessarily close to it in the run of words on a page. This is especially important for a text such as a periodical where particular words may be at some distance from one another in terms of chronological progression (for instance, words in car advertisements from one decade as compared to words in car advertisements in a different decade), but may be closer when understood in terms of the way readers experience them (for instance, the episodes of a work of serialized fiction). As we will see, some of these words also lend themselves to being georeferenced and visualized on a map, thus enabling the researcher to hold relationships of semantic similarity up against geographical similarity.

As explained above, a word embedding model allows a researcher to choose one or more words and discover the other words which cluster most closely with it in the multidimensional vector

space of the model. However, simply searching for the vectors of place names within the *WHM* proves manifestly inadequate, due to the heavy use of place names in advertisements (see for example fig. 4). Place names appear disproportionately often in lists of other places in advertisements, and the strength of this signal can overpower the signals from other contexts in which that place name appears. Searching for the vectors closest to *winnipeg*, for instance (columns 2 and 3 in figure 5), turns up many street addresses, as well as city names and bigrams such as *saskatoon_vancouver* which can be traced back to lists of cities in which an advertiser has locations. Extending the search beyond 20 terms reveals the names of companies and corporations (*brandon_distributors*, *southey_hardware*, *executive_offices*, *international_laboratories*), the volume of which further cements the link with advertising.

[INSERT FIG. 4-5 ABOUT HERE]

The vector for *canada* (columns 4 and 5 in figure 5), meanwhile, turns up neighboring vectors which predominantly refer to companies and manufacturers (*waltham_products*, *c_turnbull*, *e.b.eddy*) or other words which, upon cross-referencing with the plain text of the corpus, appear frequently in advertisements (e.g., *the only sewing machine made in canada of canadian materials by canadian workmen*). While *toronto* and *gravenhurst* are two place names that do appear in the list, the latter appears only in the text of advertisements for the Rubberset Company (*rubberset company ltd factories at toronto and gravenhurst*). The word *toronto* is in much wider use throughout the corpus, but the model shows that here again it is difficult to escape the strength of its signal within advertisements. We can also seed queries with multiple words to find the vectors closest to all those terms. Columns 6 and 7 in figure 5 give the twenty vectors closest

to a list of Canadian cities. While this query is more successful at turning up other geographical referents (note too that the similarity ratings are all higher than for the *winnipeg* and *canada* searches), as with the previous search there is plenty of evidence that the results it delivers are also connected to advertisements. The bigrams provide helpful clues in this respect: *214_second*, for instance, refers to advertisements for musical instruments from the Heintzman Company, while *halifax_moncton* comes from advertisements for the Canadian Westinghouse Company.[9]

In both single- and multiple-word queries, then, place names are overdetermined by the commercial and advertising contexts in which they so frequently appear. For the purposes of exploring the workings of spatial rhetoric within the corpus, these findings are not very useful. What they do reveal, however, is the extent to which geographical references in this corpus are embedded within advertising boilerplate, the very repetitiousness and formulaic qualities of which might be disregarded by some literary researchers but seized upon by economic historians or geographers: a salient reminder to the analyst to be attentive to the affordances of the periodical and the genres within it.[10] Plotting these advertisement-derived place names cartographically could potentially yield maps of interest for research questions around advertising and commercial enterprise in early twentieth century Canada: a geography of early twentieth-century western Canadian manufacturing, for instance, or of urban commercial activity with street-level granularity.

Exploring Similarity through Cluster Dendrograms

To probe this idea further, we can return to the list of words closest to the *canada* vector in figure 5, and, adding to the query the one other country that appears in the list (*united_states*), work

iteratively to generate a list of additional place names: *united_states, canada, great_britain, united_kingdom, america, british_isles, europe, argentine_republic, australia, mexico* and *south_africa*. This has the effect of querying the model for words which are associated both with Canada and with the United States, rather than with the advertising and commercial contexts that initially emerged, and produces a more geographically coherent list of words (fig. 6).

[INSERT FIG. 6 ABOUT HERE]

Here, at last, we do seem to have arrived at something approaching a place name vector, one of many that could have been obtained had we chosen to seed a query with a different set of place names. Names of countries with readily identifiable latitude and longitude coordinates sit alongside terms that are not so easily georeferenced: *countries*, *other_countries*, *dominions*—opening up the possibility of bringing together the domains of the geographical with the discursive. In order to better understand the relationship between these terms—and the imbricating domains they represent—we can plot them using a cluster dendrogram to find the similarities between them (fig. 7).

[INSERT FIG. 7 ABOUT HERE]

As we move upwards along the y axis looking for the terms that cluster together, we see that the closest pair are *australia* and *south_africa*: two Commonwealth countries which, as settler-invader colonies in the southern hemisphere, occupy similar geopolitical positions. The next closest pair are *east_indies* and *straits_settlements* (British territories located near the Straits of Malacca); while these are not synonyms, they are geographically overlapping categories. It is

worth noting for methodological reasons that the term *straits_settlements* appears a mere six times in the corpus, which is not enough to register in one of the topics in a topic model and which a human reader might easily pass over, so it is an example of a relationship that a word embedding model is able to identify even with relatively sparse data. The next closest similarity is *new_zealand*, which is most similar to the dyad of *australia* and *south_africa*: another Commonwealth settler-invader colony in the southern hemisphere. The next most proximate are *great_britain* and *united_states*: although very different countries (and with very different relationships to Canada), they behave similarly within the corpus: they occupy similar slots in geopolitical discussions and are also frequently joined together in the phrase *great_britain and the united_states*. As we continue moving up the y axis, other pairings emerge: *france_belgium* and *italy_france*, to which another 2-country relationship—*britain_france*—is also similar, then the pairings *russia* and *italy*, *japan* and *asia*, *mexico* and *central_america*, *scandinavian_countries* and *belgium_holland*, and *argentine* and *argentina*, relationships whose similarity does not require elaboration. The cluster diagram is thus an effective way of showing which words are semantically close to one another, and prompting the interpreter to ask why. Some clusters can be readily linked to geographical proximity (eg. the pairing *scandinavian_countries* and *belgium_holland*), but for others the relationship is something different, perhaps synecdochical (*mexico* and *central_america*; *japan* and *asia*) or a less easily specified form of similarity (*africa* and *south_america*, which given the prominence of the Commonwealth laid out above, might suggest similarity via racialized otherness). Most obviously, the cluster dendrogram suggests that in the *WHM*'s articulation of place, the British Commonwealth is an important structuring principle (as in the case of the four Commonwealth countries where English is either the dominant language or one of two dominant languages, and

also the pairing of *britain* and *overseas_dominions*, a dyad which is close to a third term, *dominions*). Other pairings give insight into the tacit assumptions that are at work when generic terms such as "countries" are used: the closest term to *countries* is *european_countries,* which as a dyad is also close to *other_countries*, meaning that a word that often occupies the same slot as *countries* is *Europe*—it is, in other words, the unmarked choice.

These are just a few of the more obvious ways in which spatial, geographical and national similarities can be drawn out from the result of word embedding model queries. Michael Gavin and Eric Gidal pursue a comparable confluence of similarities within nostalgic anti-modern discourse associated with Ossianic poetry, seeking to uncover how it corresponds to the geography of Scotland. They situate their geospatial text analysis at the intersection of two key assumptions: the distributional hypothesis, and the principle of "spatial autocorrelation," which holds that nearby places will tend to have similar characteristics at similar times (see note 7). Combining these assumptions results in the claim that similar places at similar times will tend to be described using similar terms.[11] While there is plenty of evidence—including their essay—that some aspects of the text/place interface do function in this way, I want to put some pressure on this assumption by reasserting the importance of understanding how the more obscured parts of a discursive structure are operating, such that a reader must "read between the lines" to capture the *infrequently* articulated connections that are precisely what is at risk of being missed in a distant reading approach. The challenge for word vector analysis is to be able to identify the subtle, submerged ways in which texts, and indeed entire periodicals, join places and concepts together into a spatial imaginary, keeping in view how nuances of meaning are achieved through stylistic and syntactic as well as lexical choices. While the cluster dendrograms deliver plenty of

obvious connections, there are also less easily identifiable relationships that could be further queried through the model. Taking the *africa*/*south_america* pairing, for instance, and using vector subtraction to pose the query *africa* is to *south_america* as *south_america* is to …? produces *asia*, *egypt* and *india* as the closest terms after the two search terms, suggesting that racialized otherness may indeed be part of what is causing *africa* and *south_america* to be associated.

Combining K-means Clustering and GIS Maps

So far, the queries above have been undertaken in a supervised manner (have had human input into them in the form of the search terms chosen). It is also possible to run clustering algorithms that are unsupervised (without human input at the initial stages), one of which is k-means clustering. This generates lists of the words in the most tightly clustered groupings—words whose proximity to one another is the most pronounced, as seen for example with the overlapping words *western*, *home* and *monthly* in the tSNE plot in figure 2. One such list is given in figure 8, which has been taken from a much longer list of clusters produced by the k-means clustering algorithm: one among several containing words that indicate there may be something of potential geospatial interest. Such a list can then be used as the query terms for further lists, and the resulting longer lists plotted using principal component analysis, as seen in figure 10.[12]

[INSERT FIG. 8 ABOUT HERE]

Here we can see the presence of topographical referents—*west*, *river*, *miles*—used to orient oneself on a map or to a landscape. This cluster also succeeds in doing what our earlier queries did not: mixing geographical referents from both Canada *and* elsewhere, and combining words that can be georeferenced (*assiniboine_river*, *fort_mcmurray*, *rotorua*, *orkney_islands*) with ones that cannot (*empties_into*, *watershed*, *most_northerly*, *main_line*). It suggests a semantic field that is anchored in real-world geographical locations but also goes beyond that in the way it talks about the relationship between human beings and place.

[INSERT FIG. 9 ABOUT HERE]

Figure 9, which plots the 250 words closest to the k-means cluster in figure 8, does seem to approach what we might term a Canadian spatial imaginary: while traces of advertising discourse are still present (*scenic_shasta* appears solely in advertisements for American rail travel, for instance), the predominance of non-metropolitan words signals that we have moved away from the discourse of towns, cities and provinces that proved so difficult to escape earlier. This is not to say that a "genuine" Canadian spatial imaginary will be rural and non-metropolitan. Rather, it is to propose that in the diffuse cloud of overlapping spatial imaginaries which range across the continuum between urban and rural, we have located one corner of vector space where the absence of these metropolitan words suggests a way of thinking about space that bears fewer traces of the language of advertising, and can therefore be assumed to be closer to a spatial imaginary that is derived less from advertisements and more from other kinds of copy in the magazine. I am not suggesting that advertising is something to be ignored or excised from an investigation such as this: it has its own important geospatial imaginary, but it is only one part of a much bigger picture.

[INSERT FIG. 10 ABOUT HERE]

We can explore what can be gained by juxtaposing semantic proximity to cartographic relationships by plotting these place names in a GIS. Figure 10 superimposes colored circles on clusters which are then plotted in figures 11, 12, 13 and 14 (with the exception of the red circle, which is still semantically coherent, as the words within it can all be linked to the field of railway travel). The blue, green and orange circles have been placed in an ad hoc manner, in order to see whether the geographically diverse places they include—for instance *chesterfield_inlet*, *mackenzie_river*, *athabasca_river*, *fort_mcmurray*, *fort_smith*, *fort_resolution*—have any sort of spatial relationship that could be elucidated via cartographic representation. Figure 11 shows all three clusters on a single map; to make the geographic distribution easier to discern, figures 12–14 represent each cluster on a single map.

[INSERT FIG. 11 ABOUT HERE]

What, if anything, do these different clusters suggest about the words they anchor to the map? Although it is perilous to try to find unifying labels for widely disparate places, there is some broad geographical patterning to be observed within each cluster.

[INSERT FIG. 12 ABOUT HERE]

The points in figure 12 I have posited as being related to inland waterways; though some of these are on or near the west coast, others refer to rivers and settlements on a route through the Canadian Shield from the Arctic Ocean down to the Great Lakes.

[INSERT FIG. 13 ABOUT HERE]

The points in figure 13 refer to places that are either not within Canada or else on or near its borders. For these I have suggested the label "waterways on boundaries": boundaries between Canada and the United States, or between Canada and the rest of the world. The words *atlantic* and *pacific* both appear in this cluster, and returning to the plain text of the corpus shows that often these oceans are invoked not for themselves but for what is across or beyond them, so with further cross-referencing to the *WHM* corpus it might be possible to make a case for these place names as denoting something approximating "a vision of Canada in a larger context."

[INSERT FIG. 14 ABOUT HERE]

The points represented in figure 14 also include place names on the Canadian Shield, but in addition some other remote waterways, so I have suggested the label "far north and interior remote waterways."

These imposed categories are clearly highly subjective and contestable, and they are not presented here as the key to deciphering the clusters in figure 9: they in fact bring to the fore the

severe limitations of an interpreter who has no first-hand experience of the majority of the mapped locations, something that is a recurrent challenge within work in the spatial humanities. Rather, what is compelling about the cartographic distributions they represent is that these offer *alternative ways of spatializing*—and hence understanding the potential relationships between—the words in the k-means plot. Moving back and forwards between these different kinds of visualization sets up what we might term a "spatial dialectic," the interactions of which suggest alternative ways of conceptualizing place. The maps also demonstrate that it is possible to take a geographically "noisy" corpus such as a periodical with place references frequently mentioned in advertisements, and query it in ways that reach beyond those formulaic references to something approaching a geospatial imaginary.

Conclusion: Capturing Emergence

How effective, then, are word embedding models as analytical tools for investigating the discursive workings of periodicals, especially where those discourses do not lend themselves to the binaries structuring conceptual frameworks such as gender? To deform a periodical first into the digitized remediation of plaintext files and then into a word embedding model is to ask, as Ben Schmidt does of the body of historic newspaper articles gathered in the *Chronicling America* project, "what if we could model all relationships between words as spatial ones? Or put another way: how can we reduce words into a field where they are purely defined by their relations?" Words cannot be defined purely by their relations, of course, any more than a single theoretical frame such as post-structuralism, queer theory or reader-response theory is able to account for all the workings and effects of textual objects. But I hope to have shown here that modeling these relationships in spatial terms does have significant potential for bringing to light insights and

patterns that might not otherwise emerge, especially where a corpus is too large to reasonably be read by a single researcher. We have seen, for instance, that the *WHM* imagines Canadian spaces and the world outside Canada in largely separate ways, and that the British Commonwealth is a salient category for structuring discussions of the latter. Advertisements have proved to comprise a substantial proportion of the place names in the *WHM*, necessitating methodological strategies to filter them out and serving as a reminder to periodical scholars of the ways in which we may read past the formulaic elements of advertisements without fully accounting for the semantic weight their repeated iterations may carry. And the cluster of terms delivered by the k-means cluster plot provide both resonant suggestions for conceptualizing Canadian spaces in different ways—in terms of islands, waterways, remoteness, boundaries—and a reminder of the challenges that the researcher's own cultural and geographical placedness poses when interpreting such clusters.

What I have also sought to illustrate is that despite the intangibility of their digital existence as a multidimensional entity existing only in imagined vector space, word embeddings can help to elucidate how periodicals—considered in all their specificity as print culture artefacts rather than as a set of digitized plaintext files—can give rise to instances of emergence.[13] Pointing to the diverse affordances of periodicals and the idiosyncratic readings practices these can give rise to, Sean Latham suggests that the twentieth-century magazine opened up new kinds of agentive possibilities in similar ways to the phonograph, the film, and early forms of hypertext ("Affordance and Emergence," 2). Drawing attention to the different configurations of ways readers can navigate, and make connections between, the different components of a periodical—its articles, headers, sub-heads, image captions, advertisements, tables of contents and so forth—

he argues that these interactions produce the conditions for emergence, "the creation of interconnected networks of meaning that are not only difficult to map or anticipate, but that elude stabilizing concepts like author, intention, and even textuality" (3). Attending to these contingent, unstable and idiosyncratic ways of making meaning from a periodical, moreover, is particularly important from a book historical perspective, as they help with the work of decoupling periodicals from a conceptual framework that understands them predominantly through metaphors of the printed book (3). Seen in this way, the overlapping discourse systems and domains within a periodical such as the *WHM* are not a problem but rather a core element of constituent interest.

With their ability to identify similarities—analogical, metaphorical, synecdochical and more—across the different genres and media of which middlebrow periodicals are comprised, word embedding models offer ways of glimpsing some of the higher-order patterns which emerge when we are able to incorporate the entire discursive field of a corpus via a high-dimensional model. Word embedding models, along with other computational approaches such as topic models, can function as proxies for the "divergent and individual readings" identified by McGregor and van Orden as characterizing twentieth-century periodicals across the taste spectrum. Getting closer to what these divergent and idiosyncratic readings of a periodical might have looked like would seem to be a worthwhile task, given the difficulty of accessing how early twentieth-century readers would have used *The Western Home Monthly*. Such "algorithmic reading" is, obviously, a far cry from the way actual human readers would have made use of its affordances, but this study has nonetheless been able to gesture towards the power of word embedding models to bring into the critical viewfinder aspects of a periodical and the discourses

circulating within it which would be difficult, if not impossible, to recuperate using other analytical methods.

Notes

[1] For examples of work in this area, see D. J. Bodenhamer, J. Corrigan, and T. M. Harris, *The Spatial Humanities: GIS and the Future of Humanities Scholarship* (Bloomington: Indiana University Press, 2010), David Cooper and Ian Gregory, "Mapping the English Lake District: A Literary GIS," *Transactions of the Institute of British Geographers* 36, no. 1 (2011): 89–108, and Charles B. Travis, *Abstract Machine: Humanities GIS* (Redlands, CA: Esri Press, 2015).

[2] *The Western Home Monthly* began publication in 1899, but the corpus I use here only begins in 1901.

[3] "The digital *WHM* contains 348 issues, including: two issues from each of 1901 and 1903, ten issues from 1904, and every issue between 1905 and 1932 except for January 1916, September 1919, and March 1922. A special illustrated issue titled *The 1914 War* supplements the twelve issues from 1915" (Hannah McGregor and Nicholas van Orden, "Remediation and the Development of Modernist Forms in The Western Home Monthly," in *Reading Modernism with Machines* ed. Shawna Ross and James O'Sullivan [London: Palgrave Macmillan, 2016], 135–164, 141. McGregor and van Orden's work has been influential to my thinking in this essay; following their lead, I use *WHM* in this article to distinguish the digitized object from the physical print artefact, *The Western Home Monthly*, which it remediates.

[4] To create the word embedding model, I joined together 261 text files (each containing an individual issue of the *WHM*) into a single text file. This file was then cleaned, tokenized and

lowercased using Lincoln Mullen's tokenizers package in R. After the cleaning process, I was left with a text file containing 104,026 types and 23,605,381 tokens, reduced from the original file which had around 5,133,000 types and 25,100,000 tokens. I used this to train a model with 200 dimensions, using Ben Schmidt's wordVectors package for R which implements the word2vec algorithm originally developed at Google. At the time of writing, wordVectors was not available through the CRAN official repository for R packages, but could be downloaded here. Schmidt has two detailed blog posts explaining the package which have been crucial in helping me to understand the possibilities of word vectors; these provide an excellent starting point for anyone wishing to get started with word embedding models themselves ("Vector Space Models for the Digital Humanities," *Ben's Bookworm Blog*, October 25, 2015 and "Rejecting the Gender Binary: A Vector-Space Operation," *Ben's Bookworm Blog*, October 30, 2015).

[5] Katrin Erk, "Vector Space Models of Word Meaning and Phrase Meaning: A Survey," *Language and Linguistics Compass* 6, no. 10 (2012): 635–53, 634.

[6] In the interests of moving the argument forward, I have omitted most of the technical details about how word embedding models are constructed and queried. I have given a more detailed technical account on my website here: "Word Embedding Models: A Very Short Introduction."

[7] The idea that words which have similar meanings occur in similar contexts is known within linguistics as the distributional hypothesis, and was first put forward by J. R. Firth in 1957. See Magnus Sahlgren, "The Distributional Hypothesis," *Italian Journal of Linguistics* 20, no. 1 (2008): 33–53, 33, and Stephen Clark, "Vector Space Models of Lexical Meaning," in *The Handbook of Contemporary Semantic Theory*, ed. Shalom Lappin and Chris Fox (Malden, MA: Wiley Blackwell, 2015), 493–522, 494.

[8] See Benedict Anderson, *Imagined Communities: Reflections on the Origin and Spread of Nationalism* (London: Verso, 1983).

[9] A methodological note on the bigrams is needed here. One of the options offered by tokenizers, the R package used to prepare the corpus, is the ability to join together common bigrams with an underscore so that, for instance, *new york* becomes *new_york*. Tests that I carried out on two versions of the model – one trained on the text with bigrams joined, one trained on the text without – demonstrated that for this corpus, there were sufficient analytical gains to be had from joining bigrams to make this a corpus cleaning step worth including. The gains were particularly salient for a geographically oriented analysis, as without this step, terms such as *united states* and *united kingdom* would split into their constituent parts which formed part of other phrases, and were thus more difficult to run queries on. OCR (optical character recognition) errors meant that occasionally words which had been erroneously split into two (sometimes because they fell over a line break) would reappear as bigrams (for instance *splen did* became *splen_did*). While it would of course be best to avoid OCR errors altogether, this is not feasible for most large corpora derived from digitized materials, and an example like this shows that it is preferable to have *splendid* and *splen_did* appearing as two distinct words in the corpus, rather than to have *did* appearing as a word on its own and adding noise to the corpus. Where such "OCR bigrams" appear in the figures below, it will be observed that they do in any case often cluster near the actual word, or appear in a semantically related cluster.

[10] This is not, of course, to overlook the fact that "literary" readings of the spatial imaginary constructed by the magazine would have been heavily influenced by its advertisements for both initial readers and contemporary readers.

[11] Michael Gavin and Eric Gidal, "Scotland's Poetics of Space: An Experiment in Geospatial Semantics," *Journal of Cultural Analytics*, November 16, 2017.

[12] Principal component analysis (PCA) is another technique for reducing the dimensionality of a high-dimensional dataset so that it can be represented in two or three dimensions. It does this by finding the axis along which the variance between points in the data set is maximised (represented along the PC1 axis), and then another axis orthogonal to the first axis which accounts for as much of the rest of the variance in the rest of the dataset as possible.

[13] Sean Latham, following Katherine Hayles, characterizes emergence as "the creation of meanings and behaviors generated by the multiple ways in which textons [strings of signs] can interact with one another" ("Affordance and Emergence: Magazines as New Media," paper presented at the MLA Convention, Boston, MA, January 2013, 4).